Journal of Physiotherapy 69 (2023) 100–107

Journal of PHYSIOTHERAPY

journal homepage: www.elsevier.com/locate/jphys

Research

External validation and updating of prognostic models for predicting recovery of disability in people with (sub)acute neck pain was successful: broad external validation in a new prospective cohort

Roel W Wingbermühle [a,b], Alessandro Chiarotto [b], Emiel van Trijffel [c], Martijn S Stenneberg [a,d], Ronald Kan [a], Bart W Koes [b,e], Martijn W Heymans [f]

[a] SOMT University of Physiotherapy, Amersfoort, The Netherlands; [b] Department of General Practice, Erasmus MC, University Medical Center, Rotterdam, The Netherlands; [c] Ziekenhuisgroep Twente, ZGT Academy, The Netherlands; [d] Department of Physiotherapy, Human Physiology and Anatomy, Experimental Anatomy Research Department, Vrije Universiteit Brussel, Belgium; [e] Department of Sports Science and Clinical Biomechanics, Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark; [f] Department of Epidemiology and Data Science, Amsterdam University Medical Center, VU University Medical Center, Amsterdam, The Netherlands

KEY WORDS

External validation
Neck pain
Clinical prediction model
Prognosis
Prognostic model
Recovery

ABSTRACT

Question: Can existing post-treatment prognostic models for predicting neck pain recovery (primarily in terms of disability and secondarily in terms of pain intensity and perceived improvement) be externally validated and updated at the end of the treatment period and at 6 and 12 weeks of follow-up in a new Dutch cohort of people with neck pain treated with guideline-based usual care physiotherapy? **Design**: External validation and model updating in a new prospective cohort of three previously developed prognostic models. **Participants**: People with (sub)acute neck pain and registered for primary care physiotherapy treatment. **Outcome measures**: Recovery of disability, pain intensity, and perceived recovery at 6 and 12 weeks and at the end of the treatment period. **Results**: Discriminative performance (c-statistic) of the disability model at 6 weeks was 0.73 (95% CI 0.69 to 0.77) and reasonably well calibrated after intercept recalibration. The disability model at 12 weeks and at the end of the treatment period showed discriminative c-statistic performance values of 0.69 (95% CI 0.64 to 0.73) and 0.68 (95% CI 0.63 to 0.72), respectively, and was well calibrated. Pain models and perceived recovery models did not reach acceptable performance. Cervical mobility added value to the disability models and pain catastrophising to the disability and pain models at 6 weeks. **Discussion**: Broad external validation of the disability model was successful in people with (sub)acute neck pain and clinicians may use this model in clinical practice with reasonable accuracy. Further research is required to assess the disability model's clinical impact and generalisability, and to identify additional valuable model predictors. **Registration**: https://osf.io/a6r3k/ **[Wingbermühle RW, Chiarotto A, van Trijffel E, Stenneberg MS, Kan R, Koes BW, Heymans MW (2023) External validation and updating of prognostic models for predicting recovery of disability in people with (sub)acute neck pain was successful: broad external validation in a new prospective cohort. *Journal of Physiotherapy* 69:100–107]**

Introduction

Neck pain is common and remains one of the leading causes of disability in most countries.[1,2] Its burden is likely to increase even further, warranting greater need for rehabilitation services in primary care.[3,4] Identification at intake of patients with neck pain who are unlikely to recover enables personalised care and supports the improvement of health outcomes with potential to reduce its burden. Recovery of acute neck pain mainly takes place in the first few weeks, otherwise prognosis becomes worse potentially leading to persistent pain and disability.[5,6] Prognostic factors for predicting neck pain recovery have largely been established;[7,8] however, individual factors cannot provide sufficient information to be used for accurate individualised outcome predictions.

Prognostic models provide a personalised evidence-based approach by combining multiple predictors simultaneously to estimate a patient's future individual outcomes (eg, neck pain intensity or neck pain-related disability).[9,10] Several prognostic models for neck pain have been developed; however, methodological shortcomings are common (eg, small sample size, no correction of overfitting, lack of reporting of key performance measures, and limitations in measurement of predictors and outcomes) and very few models have been externally validated.[11,12] Recently, a model for people with neck pain was developed and internally validated in a Dutch cohort of patients treated with manual therapy, predicting post-treatment recovery of disability with good discriminative performance.[13] This disability model may have good potential to inform primary care clinicians about individual prognoses of people with neck pain after

https://doi.org/10.1016/j.jphys.2023.02.002

Table 1
The three models that underwent external validation.

Model name	Model calculation	Area under the curve (95% CI)	R^2 (95% CI)
Disability model (DModel)	−2.64 + 0.28*Subacute pain + 0.88*Chronic pain + 0.11*Baseline disability + 0.02*Age + 0.28*Sleeping problem + 0.02*FABQ-PA	0.74 (0.72 to 0.75)	0.21 (0.19 to 0.23)
Pain model (PModel)	−5.94 + 0.18*Subacute pain + 0.83*Chronic pain + 0.16*Baseline pain intensity + 0.34*BNQ-AD + 0.01*Age	0.67 (0.66 to 0.69)	0.09 (0.08 to 0.11)
Perceived improvement model (PIModel)	4.54 + 0.14*Subacute pain + 0.82*Chronic pain + 0.35*Low back pain + 0.03*FABQ-PA + 0.01*Age − 0.03*Baseline disability − 0.4*Previous episode + 0.33*Sporting activities	0.67 (0.65 to 0.69)	0.09 (0.07 to 0.11)

treatment. However, a model's broad external validation is a crucial step before it can be advocated for clinical use[14] and there should be an ongoing process of model validation and updating.[15,16] Models for recovery of pain and perceived improvement have also been developed but they do not meet commonly used thresholds for discriminative performance criteria; however, they still exhibit reasonable performance and may benefit from model updating.[13]

Therefore, the research question for this study was:

> Can existing post-treatment prognostic models for predicting neck pain recovery (primarily in terms of disability and secondarily in terms of pain intensity and perceived improvement) be externally validated and updated at the end of the treatment period and at 6 and 12 weeks of follow-up in a new Dutch cohort of people with neck pain treated with guideline-based usual care physiotherapy?

Method

An external validation study of three internally validated models for recovery of neck pain was performed. This study was prospectively registered on the Open Science Framework and is reported according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) recommendations.[17] The models that were externally validated were a disability model (DModel), a pain model (PModel) and a perceived improvement model (PIModel), as shown in Table 1.

Development cohort

The models were previously developed in ANIMO, a prospective cohort study with 12-month follow-up that ran during the period 2007 to 2009, where 345 manual therapists in the Netherlands each recruited five consecutive patients aged between 18 and 80 years (total n = 1,311) who presented to physiotherapy clinics with non-specific neck pain of any duration, with or without arm pain. Participants were required to provide baseline data and sign informed consent to be deemed eligible (n = 1,193). Patients with red flags were excluded and no serious pathology was assumed. Participants received usual care manual therapy, such as specific joint mobilisations, high-velocity thrust techniques, myofascial techniques, advice or exercise. Follow-up timepoints were at the end of the treatment period and at 12-months follow-up. More detail on this cohort is provided elsewhere.[18]

Validation cohort

For external validation, data from the PRONEPA cohort, which ran from November 2020 to April 2021, with a 12-week follow-up was used. PRONEPA was a prospective cohort study (registered at https://osf.io/u8rnw/, ethics committee permission METCZ20200178) that primarily aimed to evaluate prognostic factors that predict development of chronic neck pain in people with (sub)acute neck pain (< 12 weeks), with or without radicular symptoms, registered for physiotherapy treatment. PRONEPA 2020 to 2021 included a convenience sample of 586 participants with neck pain, recruited by 102 physiotherapists who were graduating from a Master of Science program in manual therapy at SOMT University of Physiotherapy, Amersfoort (the Netherlands).

Inclusion criteria were: primary complaints of neck pain grade 1, 2 or 3 according to the Neck Pain Task Force;[19] age ≥ 18 years; and minimum 3 days to maximum 12 weeks of neck pain. Exclusion criteria were: past or current cervical fractures; congenital disorders affecting cervical functioning; systemic diseases or neurological disorders affecting cervical functioning; past or actual malignancy; and previous cervical surgery. Patients received interventions in accordance with the 'Clinical Practice Guideline for Physical Therapy Assessment and Treatment in Patients with Nonspecific Neck Pain'.[20] This included information on the benign nature of the condition, advice to stay active, neck muscle strengthening exercises and cervical spine mobilisations. No high-velocity low-amplitude interventions were applied. Participant characteristics and the models' predictors were collected by the physiotherapist at baseline and 6 weeks, and models' outcomes (Dutch version of the Neck Disability Index, Numeric Pain Rating Scale and Global Perceived Effect) at 3, 6 and 12 weeks, 6 months, and at the end of the treatment period (only for Neck Disability Index and Numeric Pain Rating Scale), which was defined by date of discharge.

Validation procedure

We described and compared case-mix differences (ie, participant characteristics data and outcome occurrence) and study characteristics (ie, recruitment period, setting, inclusion/exclusion criteria and treatment) between the development and validation cohorts, and we tested the models' performance in the validation cohort by examining discrimination, calibration and overall performance measures. The number of events in the recovered and non-recovered disability, pain and perceived improvement outcome groups for a minimum of 100 to 200 events was checked a priori at each follow-up of the validation cohort, as advised for validation studies that predict binary outcomes.[21] External validation at 6 weeks, 12 weeks and at the end of the treatment period (if < 12 weeks) was performed. We also evaluated whether the models could be updated by adding additional potential predicting variables.

Outcomes

Comparable with the derivation study, 'recovery' was used as an umbrella term for three different constructs: pain intensity, disability and perceived improvement. The outcome measure defining recovery with respect to pain was the Numeric Pain Rating Scale (11-point Likert scale), dichotomised into ≤ 2 (recovered) or > 2 (not recovered). The outcome measure defining recovery with respect to disability was the Neck Disability Index (0 to 50 scale, transformed to % by multiplying by 2) dichotomised into < 8% (recovered) or ≥ 8% (not recovered). The outcome measure defining recovery with respect to perceived improvement was Global Perceived Effect (7-point Likert scale), where the response options 'very much better' or 'much better' defined recovery.

Comparison of study characteristics

Both cohorts were recruited by students graduating from a Master of Science program in manual therapy at the same institution. Usual

care physiotherapy treatment was provided in both studies. The manual therapists in the development study had more work experience (mean work experience 19.3 years, SD 7.1) and qualified manual skills compared with the physiotherapists in the validation study (mean work experience 5.4 years, SD 4.7), who were final year Master of Science in manual therapy students. Therefore, the manual therapists may have added high-velocity thrust techniques and specific joint mobilisation techniques. It was expected that the manual therapy students were providing care according to current Dutch guidelines.[20] Both studies excluded red flags and included people aged > 18 years with non-specific neck pain with or without arm pain, and with or without trauma. The validation cohort contained 10 (1.7%) participants aged > 80 years and the development study excluded those aged > 80 years. The development study included people with neck pain of any duration, the validation study included those who had had neck pain for a minimum of 3 days to a maximum of 12 weeks. For model updating, additional history, physical examination and psychosocial variables were available in the validation cohort.

Data analysis

Missing data

We described missing predictor and outcome values and analysed the data to assume the missing data mechanism (Little's test, t-test, chi-square test and logistic regression analysis) to decide whether multiple imputation was needed.

Statistical validation of the models' performance

We tested linearity and model assumptions and compared observed outcomes with those predicted by the models in the validation cohort in terms of discrimination and calibration measures.[9] We calculated the model's linear predictors (*lp*) and individual probability of recovery for disability, pain and perceived improvement as $p\,(y = 1) = 1/(1 + e^{-lp})$ for all participants at 6-weeks and 12-weeks follow-up, and at the end of the treatment period.[22] Each model's overall performance was estimated by Nagelkerke's R^2 and Brier scores.

Discriminative performance: Discriminative performance indicates whether a model can distinguish between people with neck pain with and without recovery. It was calculated as the concordance (c) statistic, which is comparable with the area under the curve (AUC) of the receiver operating characteristic (ROC) curve for binary data.[9] Discriminative performance was a priori considered acceptable if the AUC was ≥ 0.70.[23]

Calibration performance: Calibration performance refers to the agreement between a model's predicted risks and observed outcomes.[24] We performed calibration-in-the-large and present the models' calibration slopes and calibration plots.[24] The models were re-estimated in the validation cohort using the linear predictor (*lp*) and model: logit (*y*) = *a* + *b* x *lp*.[9,24,25] Calibration was tested as deviation from the ideal calibration slope of 1 and the intercept of 0 using the model with an offset procedure. Calibration plots' probabilities were calculated to allow observation of whether all decile groups closely fit the perfect 45 deg line of identity.[9,24] Statistical analyses were performed using commercial[a] and open source[b] software.

Updating of models: We evaluated whether updating each model enhanced its performance through adjustment of the model's intercept using the calibration intercept, and the model's regression coefficients using the calibration slope.[26–28]

Additional variables were available in the validation cohort, and we tested whether a limited number of potential predictors improved the models. First, from physical examination, *cervical mobility* and *endurance of the anterior neck muscles* were used for updating the models as interventions aimed at improving these functions are effective.[29,30] Cervical mobility was measured in degrees by a total sum score of flexion, extension and both rotations using the mobile phone application[c]. Smartphone applications measuring spinal ROM are reliable and their clinical use is supported.[31] Measurement error of the Goniometro application using the CROM-device as reference appeared small.[32] Endurance of the anterior neck muscles was measured by the neck flexor endurance test;[33] however, it revealed substantial intra-rater and inter-rater reliability and a large standard error of measurement of ≥ 14.57 seconds and a minimum detectable change of 40 seconds.[34] Pain catastrophising was also considered as a potential additional predictor.[35] Catastrophising is considered a predictor for persistent pain and disability in people with chronic musculoskeletal pain and in people with whiplash-related pain.[36,37] Pain catastrophising was measured with the pain catastrophising scale, which is a reliable and valid instrument for measuring catastrophic thinking related to pain.[38,39] We evaluated whether the models improved significantly after including these potential candidate predictors ($p < 0.157$) and enhanced model performance.[15,27,28]

Table 2
Predictor variables and characteristics of participants in the validation cohort and the development study.

Predictor characteristics	Validation cohort (n = 586)	Development study (n = 1,193)
Age *(y)*, mean (SD)	44.0 (15.7)	44.7 (13.7)
Sex, n (%) female	393 (67)	823 (69)
Previous neck pain episode, n (%)	490 (84)	755 (67)
Neck pain duration, n (%)		
acute *(0 to 6 wk)*	464 (79)	420 (39)
subacute *(6 to 12 wk)*	122 (21)	138 (13)
chronic *(> 12 wk)*		513 (48)
Pain intensity (*Numerical pain rating scale, 0 to 10*[a] *and 1 to 10*[b]), median (IQR)[a] and mean (SD)[b]	6 (4 to 7)	4.8 (2.1)
Disability (*Neck Disability Index, 0 to 50)*, median (IQR)	11 (8 to 16)	12 (8 to 17)
Accompanying low back pain, n (%)	167 (29)	538 (45)
Accompanying general sleeping problems, n (%)	280 (48)	337 (28)
Fear-Avoidance Beliefs Questionnaire, Physical Activity subscale (*0 to 24*), median (IQR)	8 (4 to 13)	11 (6 to 15)
Neck Bournemouth Questionnaire, Anxiety and Depression subscale *(0 to 20)*[c], median (IQR)	5 (2 to 9)	7 (3 to 10)
Partaking in sporting activities, n (%)	404 (69)	783 (66)
Potential predictors characteristics	Validation cohort (n = 586)	
Range of motion *(deg, sum score)*, mean (SD)	62 (11)	
Neck flexor muscle endurance *(s)*, median (IQR)	30.5 (21 to 46)	
Pain Catastrophising Scale *(0 to 52)*, median (IQR)	6 (2 to 12)	

Percentages are rounded to the closest integer.
[a] Pertains to validation cohort.
[b] Pertains to derivation cohort.
[c] Sum score of 11-point numeric subscale of items 4 and 5.

Results

The predictor and outcome characteristics between the validation and derivation cohort are presented in Tables 2 and 3, respectively. Due to the difference in neck pain duration inclusion criteria, the validation cohort displayed no participants with chronic neck pain and 40.2% more participants with acute neck pain. There were no clinically meaningful differences between the other predictors (Table 2). The amount and percentage of non-recovered participants with neck pain in the validation cohort decreased from 6 to 12 weeks for all outcomes. The percentage of post-treatment non-recovered participants with neck pain between the validation and derivation study was comparable for pain intensity and differed for disability (post-treatment perceived recovery was not registered in the validation study). The number of recovered and non-recovered events in the validation cohort turned out to be between the required minimum of 100 to 200 events for all follow-up periods; for disability, the number of events exceeded 200 for all follow-up periods.

Table 3
Outcome variables in the validation cohort and development study.

Outcome variables	Validation cohort		Development study	
	(n = 586)	Not recovered n (%)	(n = 1,193)	Not recovered n (%)
6 weeks				
Pain intensity (*Numerical pain rating scale, 0 to 10*[a] *and 1 to 10*[b]), median (IQR)	2 (1 to 4)	220 (38)		
Disability (*Neck Disability Index, 0 to 50)*, median (IQR)	5 (2 to 9)	349 (60)[c]		
Perceived improvement *(Global perceived effect, 1 to 7)*, n (%)		174 (30)		
very much better	127 (22)			
much better	277 (48)			
slightly better	132 (23)			
no change	31 (5)			
slightly worse	8 (1)			
much worse	3 (1)			
very much worse	0 (0)			
12 weeks				
Pain intensity (*Numerical pain rating scale, 0 to 10*[a] *and 1 to 10*[b]), median (IQR)	1 (0 to 3)	167 (29)		
Disability (*Neck Disability Index, 0 to 50)*, median (IQR)	3 (25 to 75)	272 (47)[c]		
Perceived improvement *(Global perceived effect, 1 to 7)*, n (%)		150 (26)		
very much better	201 (35)			
much better	232 (40)			
slightly better	101 (17)			
no change	31 (6)			
slightly worse	13 (2)			
much worse	2 (0)			
very much worse	2 (0)			
End of treatment period				
Pain intensity (*Numerical pain rating scale, 0 to 10*[a] *and 1 to 10*[b]), median (IQR)	1 (0 to 2)	134 (23)	2 (1 to 2)	112 (21)
Disability (*Neck Disability Index, 0 to 50)*, median (IQR)	3 (1 to 7)	274 (48)[c]	5 (1 to 9)	290 (58)[c]
Perceived improvement *(Global perceived effect, 1 to 7)*, n (%)				107 (21)
very much better			127 (24)	
much better			287 (55)	
slightly better			83 (16)	
no change			24 (5)	
slightly worse			0 (0)	
much worse			0 (0)	
very much worse			0 (0)	

Percentages are rounded to the closest integer.
[a] Pertains to validation cohort.
[b] Pertains to derivation cohort.
[c] 0 to 50 value transformed to % by multiplying by 2 and dichotomised into $<$ 8% (recovered) or $\geq$ 8% (not recovered).

Model performance

Missing data

The number of variables with missing data and the amount of missing data was very low, with the vast majority between 0 and 1%. Six variables had marginally more than 1% missing values: the cervical mobility predictor 1.2%, the three outcome variables at 6 weeks 1.4%, post-treatment pain intensity 1.2%, and post-treatment disability 1.5%. There was little gain from multiple imputation for these low proportions of missing data.[40] In addition, after analysing the missing data and checking the researcher's logbooks for reasons for missing outcomes, it was assumed that the missing completely at random missingness was plausible. Consequently, it was decided that there was no need for multiple imputation and complete case analysis was acceptable.

Model performance measures before updating

We tested and evaluated linearity and concluded that non-linear transformation would not be advantageous.[41] The models' validation performance is described in Table 4. The disability model at 6 weeks showed acceptable discriminative performance with a c-statistic equal to 0.73 (95% CI 0.69 to 0.77). The disability models at 12 weeks and at the end of treatment showed discriminative performance values of 0.69 (95% CI 0.64 to 0.73) and 0.68 (95% CI 0.63 to 0.72), respectively. The pain models and perceived improvement models did not reach acceptable levels of the performance measures

Table 4
Performance of the three models before updating.

Model name	Discrimination (*c-statistic)* logit (95% CI)	Nagelkerke's R^2	Brier Score	Testing calibration	
				in-the-large *(intercept)*	*(slope)*
Disability model (DModel)					
6 weeks	0.73 (0.69 to 0.77)	0.20	0.20	0.60[a]	1.10
12 weeks	0.69 (0.64 to 0.73)	0.14	0.22	−0.06	0.83[a,b]
post-treatment[c]	0.68 (0.63 to 0.72)	0.12	0.23	−0.05	0.76[a,b]
Pain model (PModel)					
6 weeks	0.66 (0.62 to 0.71)	0.10	0.22	0.29	0.32[a]
12 weeks	0.66 (0.61 to 0.71)	0.09	0.19	−0.16	0.31[a]
post-treatment[c]	0.61 (0.56 to 0.67)	0.04	0.17	−0.69[a]	0.21[a]
Perceived improvement model (PIModel)					
6 weeks	0.53 (0.48 to 0.58)	0.00	0.21	−1.91	0.21
12 weeks	0.54 (0.48 to 0.59)	0.00	0.19	−2.59	0.31

[a] Significant deviation (intercept from 0, slope from 1) for test *lp* fit.
[b] Not significant deviation for test intercept and slope separate with offset procedure.
[c] If $<$ 12 weeks.

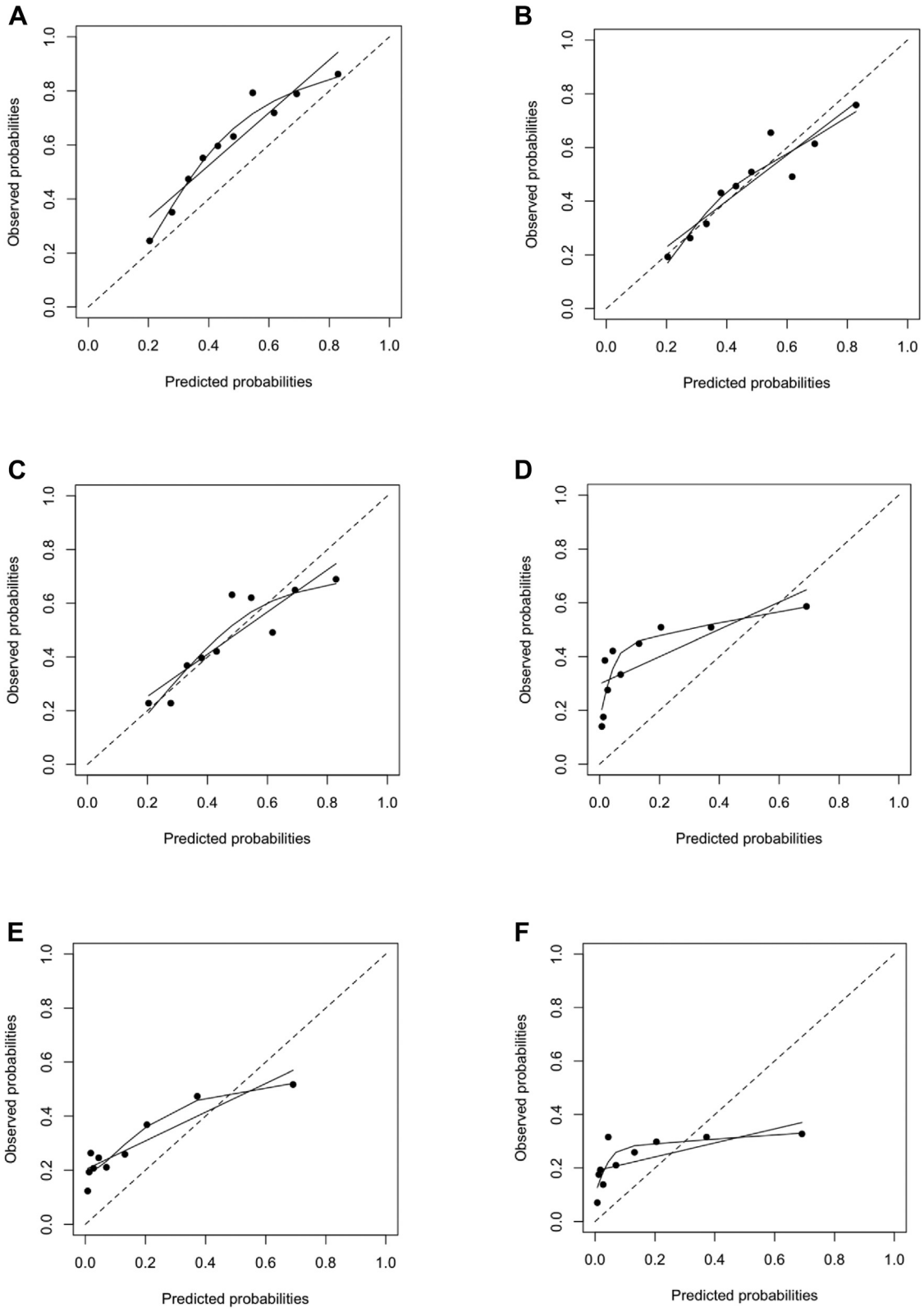

Figure 1. Calibration curves of the three models at validation. Disability model calibration curve at 6 weeks (A), 12 weeks (B) and the end of treatment (C). Pain model calibration curve at 6 weeks (D), 12 weeks (E) and the end of treatment (F). Perceived improvement model calibration curve at 6 weeks (G) and 12 weeks (H).

(Table 4). The pain models showed discriminative performance values of 0.66 (95% CI 0.62 to 0.71), 0.66 (95% CI 0.61 to 0.71) and 0.61 (95% CI 0.56 to 0.67) at 6 weeks, 12 weeks and at the end of treatment, respectively. The perceived improvement models showed discriminative performance values of 0.53 (95% CI 0.48 to 0.58) and 0.54 (95% CI 0.48 to 0.59) at 6 weeks and 12 weeks, respectively. Calibration curves are displayed in Figure 1.

Model updating

We assessed model updating for the disability model and pain model. Based on the model performance, it was decided that further testing of the perceived improvement model was not useful. If intercept and/or slope values differed significantly after testing with the logit (y) = a + b x lp offset procedure, the models were updated

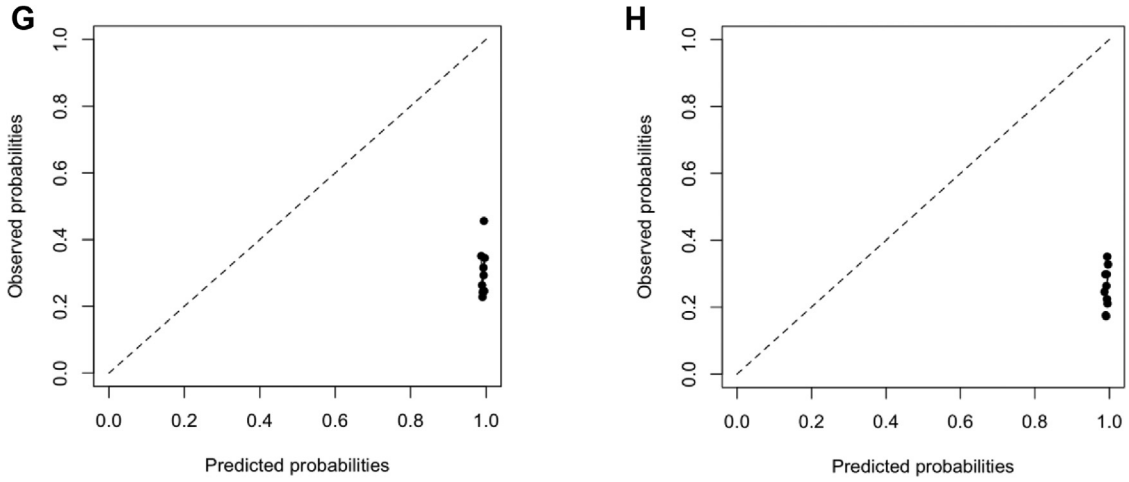

Figure 1. Continued.

with the values found, and the models' performance was subsequently re-evaluated. The calibration performance of the disability model at 6 weeks clearly improved from intercept correction using the found 0.6 value (Figure 2A); the discriminative performance did not change after this correction and remained at the same acceptable performance of c-statistic of 0.73 (95% CI 0.69 to 0.77). The other models' calibration performance did not improve, and discrimination remained identical. For further testing, the recalibrated 6-week disability model was used and the other models were not adjusted.

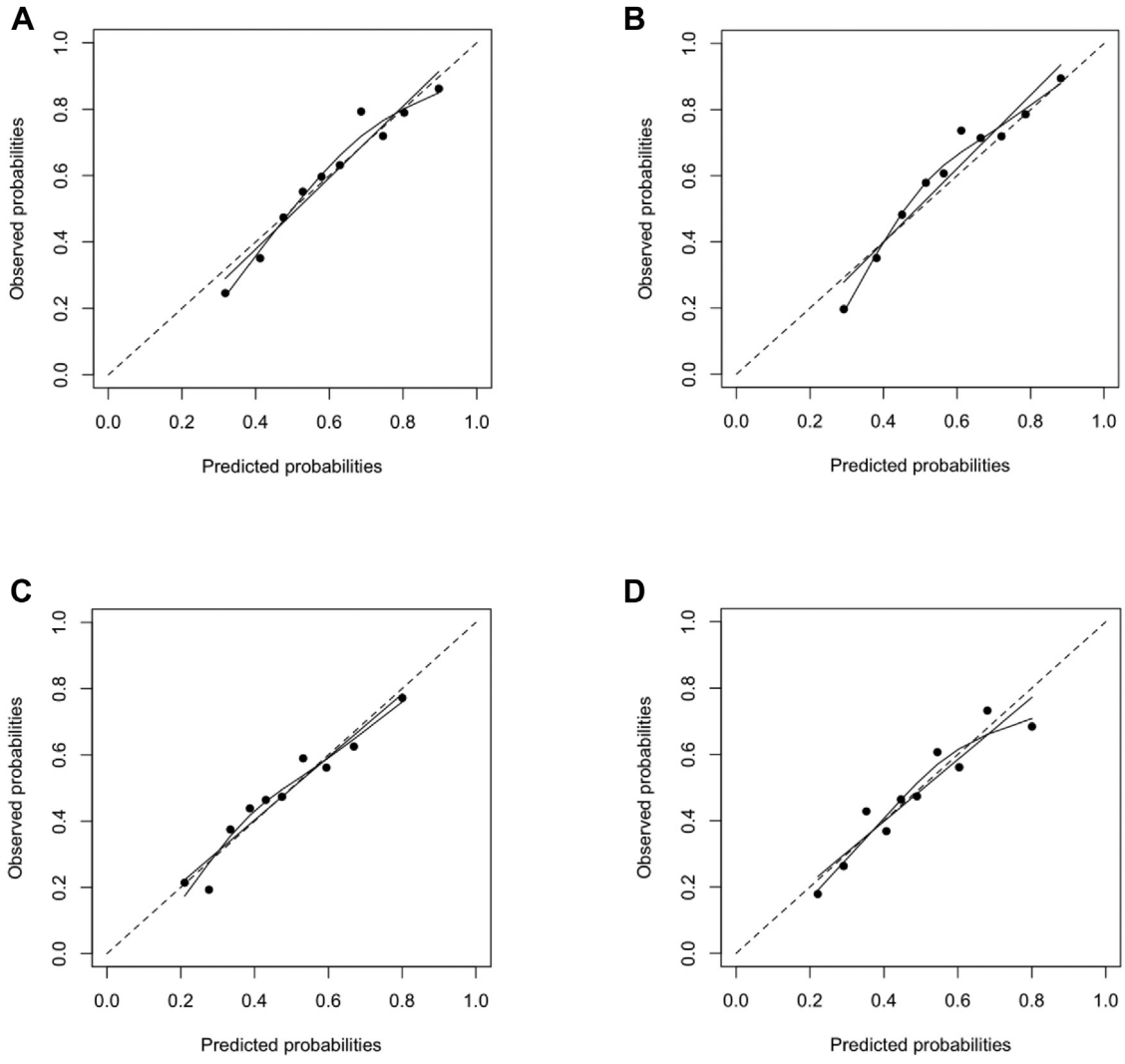

Figure 2. Calibration curves of the disability model after updating.
Adjusted disability model calibration curve at 6 weeks, after recalibration with the 0.6 calibration intercept (A), after adding the cervical mobility and pain catastrophising predictors and recalibration with the intercept (B). Disability model calibration curve at 12 weeks, after adding the cervical mobility predictor and recalibration with the intercept (C). Disability model calibration curve at the end of treatment, after adding the cervical mobility predictor and recalibration with the intercept (D).

Testing the additional variables revealed that the cervical mobility and pain catastrophising variables added significantly ($p < 0.157$) to the 6-week recalibrated disability model. The cervical mobility variable added significantly to the 12-week and post-treatment disability models. The pain catastrophising variable added significantly ($p < 0.157$) to the 6-week pain model. The neck flexor endurance test variable showed no additional significant value for the models.

The significantly adding variables and their weights were included in the disability models at the three follow-ups and to the pain model at 6 weeks, and each model's performance was then re-evaluated. Adding cervical mobility and pain catastrophising variables to the 6-week recalibrated disability model slightly improved discrimination to 0.74 (95% CI 0.70 to 0.78). Adding cervical mobility to the 12-week and post-treatment disability models showed c-statistics of 0.69 (95% CI 0.65 to 0.73) and 0.69 (95% CI 0.65 to 0.73), respectively. Calibration performance of all the disability models was initially overfitted and recovered after intercept recalibration (Figure 2B, C, D). The discrimination and calibration performance of the 6-week pain model did not improve.

Discussion

The disability model for prediction of neck pain recovery remained discriminatory at 6 weeks in a different, external cohort of people with neck pain, coming from an independent physiotherapy setting with a different case-mix. At 12 weeks and at the end of treatment, it showed nearly acceptable performance: c-statistics of 0.69 (95% CI 0.64 to 0.73) and 0.68 (95% CI 0.63 to 0.72), respectively. The pain model and perceived improvement model could not be externally validated, which was expected since internal validation was also not acceptable.[13] Cervical mobility added value to the disability model at all follow-up periods and pain catastrophising also to the 6-week pain model. Model updating hardly affected discriminative and overall performance, whereas the different levels of updating were reflected in the shape of the calibration curves. The additional predictors improved model performance minimally and may have insufficient gain to be used clinically for purely prognostic purpose.

Few prognostic models for recovery of non-specific neck pain have been exposed to external validation.[11,42] Until now, no non-specific neck pain model has been successfully externally validated with reporting of both discrimination and calibration performance measures as recommended by TRIPOD.[43] One model stood out as it was evaluated in several external validation studies, whereby all these studies reported AUC values < 0.70.[44–46]

The strength of this broad external validation study is that it was conducted in a cohort with sufficient power with very few missing values. A model is more challenged in broad than narrow external validation, indicating a better test for its generalisability.[47] This disability model keeps performing well at different follow-up periods in a cohort of people with neck pain with a different case-mix, especially regarding the duration of pre-existing neck pain complaints. In addition, participants were treated recently, reflecting current physiotherapy guidelines.

The 6-week disability model needed to be recalibrated, which is often needed in validation studies and indicates a difference in baseline risk between development and validation study that was not reflected by the model predictors. This could be explained by the difference in non-recovery percentage for disability.[26,28,48] Looking at the disability models' calibration slopes revealed that some group mean values were still somewhat scattered around the perfect line of identity. This scattering may have been less if we had decided that non-linear transformation was advantageous, at the expense of clinical manageability. Furthermore, use of predictor weights gained by fitting the models anew may have improved the model's predictive performance. However, this implies model revision and subsequent external validation and is not preferred over simple recalibration.[28]

Further research is recommended to assess the disability model's clinical utility and clinical impact. Additional external validation studies in another clinical context (eg, other countries, other healthcare providers and other settings) may add knowledge to the model's generalisability. Furthermore, the model's relatively low explained variance indicates that predictors for non-recovery are still missing and the quest for additional valuable predictors continues. Additionally, it may be of interest to further evaluate the cervical mobility and pain catastrophising predictors. Although they showed minimal impact on prognostic performance in this study, being modifiable factors, they may have predictive capacity depending on specific treatments. For instance, the cervical mobility predictor may show predictive capacity depending on mobilisation treatment, and the prognostic effect of the pain catastrophising predictor may depend on cognitive-behavioural therapy.

Broad external validation of the disability model was successful, and this model is generalisable to current physiotherapy settings and can be used in clinical practice with reasonable confidence. We advocate that physiotherapists use the disability model at intake for the prognosis of people with neck pain to assist in clinical decisions concerning recovery of neck pain disability at 6 weeks. Further research is required to assess the disability model's clinical impact and generalisability.

What is already known on this topic: Clinical use of currently published models for predicting recovery of non-specific neck pain cannot be advised.
What this study adds: A model for predicting recovery of disability at 6 weeks in people with neck pain was broadly externally validated and is advised for use in clinical practice. For clinical use, the disability model's regression formula was transformed into a web-based risk calculator, which can be accessed at www.somt.nl/research

Footnotes: [a] BM-SPSS Statistics version 27.0, SPSS Inc., Chicago, USA.
[b] R software version 2021.09.01, R Core Team, Vienna, Austria.
[c] Goniometro, Human Computer Technology, Madrid, Spain.
eAddenda: Nil.
Ethics approval: The METCZ20200178 Ethics Committee(s) approved this study. All participants gave written informed consent before data collection began.
Competing interests: The authors declare that there are no competing interests.
***Source(s) of* support**: The cohort study was supported by SOMT University of Physiotherapy.
Acknowledgements: Nil.
Provenance: Not invited. Peer reviewed.
Correspondence: Roel W Wingbermühle, SOMT University of Physiotherapy, The Netherlands. Email: r.wingbermuhle@somt.nl

References

1. Vos T, Allen C, Arora M, Barber RM, Bhutta ZA, Brown A, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet*. 2016;388(10053):1545–1602. https://doi.org/10.1016/S0140-6736(16)31678-6
2. Hay SI, Abajobir AA, Abate KH, Abbafati C, Abbas KM, Abd-Allah F, et al. Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet*. 2017;390(10100):1260–1344. https://doi.org/10.1016/S0140-6736(17)32130-X
3. Safiri S, Kolahi AA, Hoy D, Smith E, Bettampadi D, Mansournia MA, et al. Global, regional, and national burden of neck pain in the general population, 1990–2017: Systematic analysis of the Global Burden of Disease Study 2017. *BMJ*. 2020:368. https://doi.org/10.1136/bmj.m791
4. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet*. 2020;396(10267):2006–2017. https://doi.org/10.1016/S0140-6736(20)32340-0
5. Hush JM, Lin CC, Michaleff Z, Verhagen A, Refshauge KM. Prognosis of acute idiopathic neck pain is poor: a systematic review and meta-analysis. *Arch Phys Med Rehabil*. 2011;92:824–829. https://doi.org/10.1016/j.apmr.2010.12.025
6. Vasseljen O, Woodhouse A, Bjørngaard JH, Leivseth L. Natural course of acute neck and low back pain in the general population: The HUNT study. *Pain*. 2013;154:1237–1244. https://doi.org/10.1016/j.pain.2013.03.032

7. Verwoerd M, Wittink H, Maissan F, de Raaij E, Smeets RJEM. Prognostic factors for persistent pain after a first episode of nonspecific idiopathic, non-traumatic neck pain: A systematic review. *Musculoskelet Sci Pract*. 2019;42:13–37. https://doi.org/10.1016/j.msksp.2019.03.009
8. Artus M, Campbell P, Mallen CD, Dunn KM, van der Windt DAW. Generic prognostic factors for musculoskeletal pain in primary care: a systematic review. *BMJ Open*. 2017;7, e012901. https://doi.org/10.1136/bmjopen-2016-012901
9. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. 2nd ed. Cham: Springer Science and Business Media; 2019.
10. Riley RD, Van Der Windt DA, Croft P, Moons KGM. *Prognosis Research in Healthcare*. 1st ed. Oxford: Oxford University Press; 2019.
11. Wingbermühle RW, van Trijffel E, Nelissen PM, Koes B, Verhagen AP. Few promising multivariable prognostic models exist for recovery of people with non-specific neck pain in musculoskeletal primary care: a systematic review. *J Physiother*. 2018;64:16–23. https://doi.org/10.1016/j.jphys.2017.11.013
12. Wingbermühle RW, Chiarotto A, Koes B, Heymans MW, van Trijffel E. Challenges and solutions in prognostic prediction models in spinal disorders. *J Clin Epidemiol*. 2021;132:125–130. https://doi.org/10.1016/j.jclinepi.2020.12.017
13. Wingbermühle RW, Chiarotto A, van Trijffel E, Koes B, Verhagen AP, Heymans MW. Development and internal validation of prognostic models for recovery in patients with non-specific neck pain presenting in primary care. *Physiotherapy*. 2021;113:61–72. https://doi.org/10.1016/j.physio.2021.05.011
14. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016;69:245–247. https://doi.org/10.1016/j.jclinepi.2015.04.005
15. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Med*. 2013;10, e1001381. https://doi.org/10.1371/journal.pmed.1001381
16. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol*. 2015;68:279–289. https://doi.org/10.1016/j.jclinepi.2014.06.018
17. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. *Eur Urol*. 2015;67:1142–1151. https://doi.org/10.1016/j.eururo.2014.11.025
18. Peters R, Mutsaers B, Verhagen AP, Koes BW, Pool-Goudzwaard AL. Prospective cohort study of patients with neck pain in a manual therapy setting: design and baseline measures. *J Manipulative Physiol Ther*. 2019;42:471–479. https://doi.org/10.1016/j.jmpt.2019.07.001
19. Guzman J, Hurwitz EL, Carroll LJ, Haldeman S, Côté P, Carragee EJ, et al. A new conceptual model of neck pain. linking onset, course, and care: The Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. *J Manipulative Physiol Ther*. 2009;32:17–28. https://doi.org/10.1016/j.jmpt.2008.11.007
20. Bier JD, Scholten-Peeters WGM, Staal JB, Pool J, van Tulder MW, Beekman E, et al. Clinical practice guideline for physical therapy assessment and treatment in patients with nonspecific neck pain. *Phys Ther*. 2018;98:162–171. https://doi.org/10.1093/ptj/pzx118
21. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: A resampling study. *Stat Med*. 2016;35:214–226. https://doi.org/10.1002/sim.6787
22. Vergouwe Y, Moons KGM, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol*. 2010;172:971–980. https://doi.org/10.1093/aje/kwq223
23. Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. 3rd ed. Wiley; 2013.
24. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014;35(29):1925–1931. https://doi.org/10.1093/eurheartj/ehu207
25. Vergouwe Y, Royston P, Moons KGM, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. *J Clin Epidemiol*. 2010;63:205–214. https://doi.org/10.1016/j.jclinepi.2009.03.017
26. Vergouwe Y, Nieboer D, Oostenbrink R, Debray TP, Murray GD, Kattan MW, et al. A closed testing procedure to select an appropriate method for updating prediction models. *Stat Med*. 2017;36(28):4529–4539. https://doi.org/10.1002/sim.7179
27. Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*. 2012;98:691–698. https://doi.org/10.1136/heartjnl-2011-301247
28. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol*. 2008;61:76–86. https://doi.org/10.1016/j.jclinepi.2007.04.018
29. Coulter ID, Crawford C, Vernon H, Hurwitz EL, Khorsan R, Booth MS, et al. Manipulation and mobilization for treating chronic nonspecific neck pain: a systematic review and meta-analysis for an appropriateness panel. *Pain Physician*. 2019;22:E55–E70.
30. Martin-Gomez C, Sestelo-Diaz R, Carrillo-Sanjuan V, Navarro-Santana MJ, Bardon-Romero J, Plaza-Manzano G. Motor control using cranio-cervical flexion exercises versus other treatments for non-specific chronic neck pain: a systematic review and meta-analysis. *Musculoskelet Sci Pract*. 2019;42:52–59. https://doi.org/10.1016/j.msksp.2019.04.010
31. Nuhmani S, Khan MH, Kachanathu SJ, Bari MA, Abualait TS, Muaidi QI. Reliability and validity of smartphone applications to measure the spinal range of motion: asystematic review. *Expert Rev Med Devices*. 2021;18:893–901. https://doi.org/10.1080/17434440.2021.1962290
32. Verstelle K. *Reliability of the Cervical Flexion Rotation Test and the C0-C2 Axial Rotation Test in an Elderly Patient Population. Master Thesis*. Free University Brussels; 2021.
33. Harris KD, Heer DM, Roy TC, Santos DM, Whitman JM, Wainner RS. Reliability of a measurement of neck flexor muscle endurance. *Phys Ther*. 2005;85:1349–1355. https://doi.org/10.1093/ptj/85.12.1349
34. Juul T, Langberg H, Enoch F, Søgaard K. The intra- and inter-rater reliability of five clinical muscle performance tests in patients with and without neck pain. *BMC Musculoskelet Disord*. 2013;14:339. https://doi.org/10.1186/1471-2474-14-339
35. Thompson DP, Urmston M, Oldham JA, Woby SR. The association between cognitive factors, pain and disability in patients with idiopathic chronic neck pain. *Disabil Rehabil*. 2010;32:1758–1767. https://doi.org/10.3109/09638281003734342
36. Walton DM, MacDermid JC, Giorgianni AA, Mascarenhas JC, West SC, Zammit CA. Risk factors for persistent problems following acute whiplash injury: update of a systematic review and meta-analysis. *J Orthop Sports Phys Ther*. 2013;43:31–43. https://doi.org/10.2519/jospt.2013.4507
37. Martinez-Calderon J, Jensen MP, Morales-Asencio JM, Luque-Suarez A. Pain catastrophizing and function in individuals with chronic musculoskeletal pain. *Clin J Pain*. 2019;35:279–293. https://doi.org/10.1097/AJP.0000000000000676
38. Sullivan MJL, Bishop SR, Pivik J. The Pain Catastrophizing Scale: Development and validation. *Psychol Assess*. 1995;7:524–532. https://doi.org/10.1037/1040-3590.7.4.524
39. Osman A, Barrios FX, Gutierrez PM, Kopper BA, Merrifield T, Grittmann L. The pain catastrophizing scale: Further psychometric evaluation with adult samples. *J Behav Med*. 2000;23:351–365. https://doi.org/10.1023/A:1005548801037
40. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. *J Clin Epidemiol*. 2019;110:63–73. https://doi.org/10.1016/j.jclinepi.2019.02.016
41. Sauerbrei W, Perperoglou A, Schmid M, Abrahamowicz M, Becher H, Binder H, et al. State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues. *Diagn Progn Res*. 2020;4:3. https://doi.org/10.1186/s41512-020-00074-3
42. Kelly J, Ritchie C, Sterling M. Clinical prediction rules for prognosis and treatment prescription in neck pain: A systematic review. *Musculoskelet Sci Pract*. 2017;27:155–164. https://doi.org/10.1016/j.math.2016.10.066
43. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med*. 2015;162:55. https://doi.org/10.7326/M14-0697
44. Schellingerhout JM, Heymans MW, Verhagen AP, Lewis M, de Vet HCW, Koes BW. Prognosis of patients with nonspecific neck pain: development and external validation of a prediction rule for persistence of complaints. *Spine*. 2010;35:E827–E835. https://doi.org/10.1097/BRS.0b013e3181d85ad5
45. Wingbermühle RW, Heymans MW, van Trijffel E, Chiarotto A, Koes B, Verhagen AP. External validation of prognostic models for recovery in patients with neck pain. *Braz J Phys Ther*. 2021;25:775–784. https://doi.org/10.1016/j.bjpt.2021.06.001
46. Myhrvold BL, Kongsted A, Irgens P, Robinson HS, Thoresen M, Vøllestad NK. Broad external validation and update of a prediction model for persistent neck pain after 12 weeks. *Spine*. 2019;44:E1298–E1310. https://doi.org/10.1097/BRS.0000000000003144
47. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016;69:245–247. https://doi.org/10.1016/j.jclinepi.2015.04.005
48. Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW. Calibration: The Achilles heel of predictive analytics. *BMC Med*. 2019;17:1–7. https://doi.org/10.1186/s12916-019-1466-7